Choudhury, Sharmin and Raymond, Kerry and Higgs, Peter L. (2008) ***A rule-based metric for calculating semantic relatedness score for the motion picture industry***. In: Workshop on Natural Language Processing and Ontology Engineering, 9th December 2008, Sydney, Australia.

A Rule-based Metric For Calculating Semantic Relatedness Score for the Motion Picture Industry

Sharmin Choudhury (Tinni) Kerry Raymond Peter Higgs
Queensland University of Technology
t.choudhury@qut.edu.au k.raymond@qut.edu.au p.higgs@qut.edu.au

Abstract

Semantic relatedness is an important part of developing ontology based systems, as it provides the ability to measure the relatedness of two concepts. This is important not only for computational linguistics but also for intelligent querying and data classification. In this paper we present a rule-based calculation for a semantic relatedness score. This metric was specifically developed for Loculus, an ontology we have developed for the Motion Picture Industry. The Metric is evaluated using three axes: the Temporal Axis, the Inheritance Axis and the Linkage Axis. The metric has been tested on pairs of concept; the resulting scores are consistent with human expectations.

1. Introduction

The Motion Picture Industry is currently undergoing a period of great change due to increasing costs, diminishing box office returns, the advent of the internet and shift from analogue to digital processes: both in terms of using digital tools, e.g. digital cameras and editing suites, as well as moving to semi-automated workflow systems [1, 2]. The increasing costs and diminishing returns demand more efficiency in the process of developing of Motion Picture, as well as the exploration of new sources of revenue through the exploitation of not just the Motion Picture but the contextual information that surrounds it. Indeed the contextual information, that can take the form of scripts, music as well as production related data, can often just as valuable and sometimes more so than the Motion Picture itself [3]. To enable the Motion Picture Industry to take full advantage of their increasingly digital processes our project aims to construct an information management system that embodies the semantics of the Motion Picture Industry, through the Loculus ontology. The ontology will then be exploited to support reuse and repurposing of data and to facilitate improvements to the production processes.

The *Loculus* ontology is designed to cover the entire production process of a Motion Picture. This differs from existing ontologies which are focused on the finished product (the Motion Picture), which is generally treated as just another multimedia object [4-6]. We intend to use the ontology to implement more intelligent data classification, query interpretation and query response within our broader system. However, before the ontology can be used in this manner, we must have a way of determining semantic relatedness of concepts within the ontology. Therefore, in this paper we present a rule-based measure of semantic relatedness score. Our measure is based upon three axes: the *Temporal Axis*, the *Inheritance Axis* and the *Linkage Axis*. These three axes yield two scores, the *Temporal Score* that comes from the *Temporal Axis* and the *Reach Score* that comes from combining the distance travelled along the *Inheritance Axis* and the *Linkage Axis* to reach one concept from another concept. The combination of the *Reach Score* and the *Temporal Score* then yields the overall semantic relatedness score of a pair of concepts in the ontology.

In the subsequent sections, we will first explore related works, before introducing the domain and going into more depth in terms of ontology. We will then present the metric and a brief discussion of the results to date.

2. Related Works

The need to determine the degree of semantic similarity, or, more generally, relatedness, between two concepts is pervasive, especially in the lexical context for ontologies such as WordNet [7]. A number of semantic measures have been proposed to evaluate the semantic link between two concepts or two groups of concepts from two different ontologies or inside an ontology [8]. Of these measures, the path-based measures proposed by Hirst [9], Leacock [10] and especially Rada's edge counting method [11] were of particular interest to us because they were designed around ontologies with tree based structures, which is similar to the structure of Loculus. In these works, the underlying principle is that the semantic relation of two concepts can be determined by calculating the semantic distance between them.

However, simple edge counting is not necessarily an accurate measure of semantic distance. Li [12] highlights the dangers of not taking into account higher level abstraction as well as the pitfalls of non-weighted edge counting.

3. The Domain

The Motion Picture Domain at its heart is concerned with the creation of the Motion Picture itself, which can take the form of feature film, short film, documentaries, animation etc. The process by which the Motion Picture is created is referred to as the production cycle. The production cycle is broken into three phases: pre-production, production and post-production. Pre-production starts in earnest in preparation for the production phase of the cycle when sufficient budget is in place. The production phase starts on the first day of shooting and ends on the last day of shooting. As soon as production ends, post production starts. Post-production encompasses everything past production and is essentially open ended. The reason for this is that while the Motion Picture exists there is always something to do, whether it be to produce the final cut, to market the final cut or to digitally re-master the Motion Picture for a new generation or simply to preserve it.

Within the industry the Motion Picture itself is also viewed as having life stages. We have identified these life stages as conception, production, utilization and destruction. Utilization in turn comprises of distribution, discovery, access, reuse/repurpose and preservation. The life stages do not map neatly over the production cycle; nor do all Motion Pictures reach all life stages. A Motion Picture is in the conception stage when it is conceived and is being fleshed out. The later part of conception would correlate with pre-production. A Motion Picture is in production life stage when it is being produced; so the later parts of pre-production, all of production and the post-production activities that end with the creation of the final cut would correlate with this stage. Utilization spans the remainder of post-production. While the Production Cycle and the Motion Picture Life Stage are related they are not the same. An easy way to distinguish between them is that the Production Cycle creates the Motion Picture, while the Life Stages define the various forms the Motion Pictures takes just before the start and during the cycle. Graphically the relationship between the two is showed in Figure 1. The Production Life Cycle and the Motion Picture Life Stages combine to give a temporal context to all activities in the Motion Picture Industry.

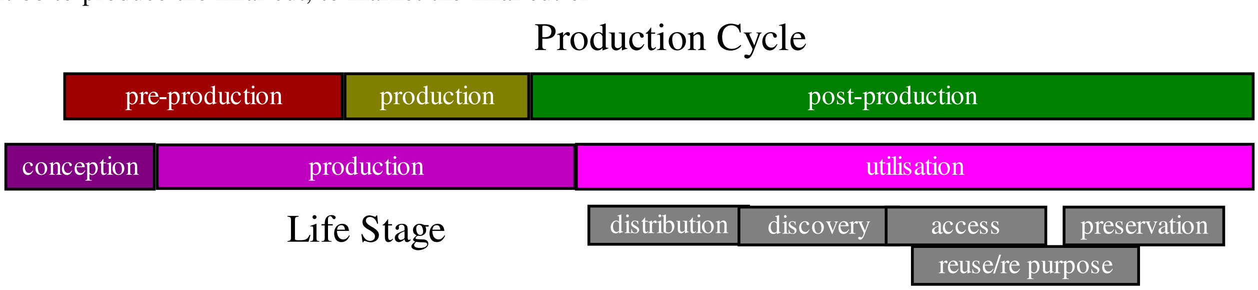

Figure 1: Relationship of Production Cycle with the Motion Picture Life Stage

4. Loculus Ontology

Loculus ontology consists of three sub-ontologies. The first is the Motion Picture Industry Terminology and Concepts ontology (MPI ontology) that models all Motion Picture specific terminology, e.g. Editing, Acting. The MPI ontology also contain artifact concept such as “script”, “music score” and the “film” itself which are actual physical artifacts that result from the production process. The second is the Agent ontology that models all Motion Picture specific agents, e.g. Actor, Editor. The third sub-ontology is the Common concepts ontology, which models common concepts that are frequently used within the Motion Picture Industry, e.g. telephone, catering. The common concepts ontology also includes common concepts that are the parents of more specific Motion Picture concepts, e.g. Action (a common concept) is the parent of Acting and Editing (which are Motion Picture Industry concepts). Of these, the MPI ontology and the Agent ontology are to be developed comprehensively. However the Common ontology is only to be developed sufficient to support the other two ontologies specific to the Motion Picture Industry.

The three sub-ontologies function together through vertical inheritance and horizontal relationship linkages, as illustrated in Figure 2 with the concept of Editor. For example, as mentioned before the MPI ontology concepts of Acting and Editing share the common concept parent of Action through a vertical inheritance link. Indeed, at the highest level of abstraction are a set of common concepts that form the root of more specific MPI concepts. The common root concepts include Action, Technique, Tool, Process, Description, Artifact and Agent. These root concepts are linked together through the concept of Motion Picture itself.

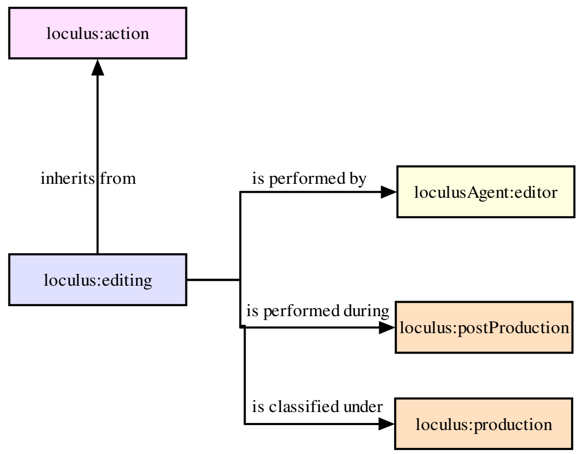

Figure 2: Ontology concept – Editing

Horizontally, Acting has a "is performed by" relationship to Actor and Editor has a "is performed by" relationship to Editor, where both Actor and Editor belong to the Agent ontology. A relationship (horizontal link) can be either weak or strong. A weak link represents a weak relationship which is non-specific and vague. An example of this is Blocking "involves agent" Actor, where "involves agent" is the weak link. Blocking is a rehearsal technique by which a scene is finalized (e.g. placement of lights, movement of actors etc) before it is shot. An Actor is one of many agents who are involved in the process. Precisely what an Actor does during blocking is hard to capture and will differ from scene to scene, Actor to Actor and indeed film to film. On the other hand, a strong link represents a precisely defined relationship that alters little between different instances of the two concepts. An example, Blocking "is performed by" a Director, where the Director's responsibilities during Blocking is well defined (he issues the instructions to the other agents involved) and therefore a strong link exists between the Director and Blocking. The weak and strong horizontal links are also used to establish perspective; not all agents have the same view of the MPI concepts, nor can all concepts in the MPI ontology be grouped in the same manner. For example, while both the Director and the Actor are involved with the process of Blocking, it is unlikely they would have the same perspective on Blocking. Similarly, even though both Acting and Editing are forms of Action, grouping according to parent class is not necessarily the most contextually appropriate grouping of the two concepts.

In addition, all concepts within the ontology have a link to one or more phases of the production cycle. Concepts can also have a link generally to the production cycle itself. This is because all concepts in the industry exist within the temporal context of the production cycle and its phases. Except for agents, the concepts within the ontology also have links to the stages of the life cycle. There are no set rules to determine the relationship between a concept and a life cycle stage. Loculus simply captures how the industry practitioners link the concepts. From extensive discussions with our industry collaborators, The Australian Film Television and Radio School (AFTRS) [13], it appears that generally the link is based on use, i.e. at what life stage is a concept most used. Agents do not have explicit life stage links. There are two reasons for this, firstly within the industry agents are not linked to the life stage explicitly and secondly because many agents who are involved with the later stages of the Motion Pictures life: access, preservation etc, are not necessarily Motion Picture industry agents but agents from the common world too numerous to be accounted for.

5. The Metric

To develop the metric for the semantic relatedness for the Motion Picture Industry, we had to take into account the nature of the industry as well as the structure of Loculus. This led to a metric that is calculated along three axes. The first axis is a *Temporal Axis* that takes into account the production cycle and Motion Picture life stage, the combination of which yields the *Temporal Score*. The second and third axes are the *Inheritance Axis* and the *Linkage axis*, which together determine the distance between two concepts in terms of the ontology: thus yielding the *Reach Score*. As the *Loculus* ontology can be thought of as a series of trees with interlinked branches, the overall *Reach Score* is determined by how far up/down and across has to be travelled from concept 1 to reach concept 2. The overall relatedness of two concepts is therefore the combined total of the scores received on the *Temporal Axis (*the *Temporal Score*) and the *Reach Score* as determined by the *Inheritance Axis* and the *Linkage Axis*. The calculation rules are given below.

- Temporal Axis
  - Production Life Cycle
    - Same cycle => + 1
    - Adjacent cycles => + 3
    - Non-adjacent cycles => +5
    - Production cycle as a whole => + 1
  - Motion Picture Life Stage
    - Same stage => +1
    - Adjacent stage => + 3

- Non-adjacent stage => +5
- If one of the concept being compared is an agent this axis is set to => 0

❖ Inheritance axis (Vertical relationships)

- First move up or down => +1
- Every move up the tree after first => +2
- Every move down the tree after first => +2
- Reaching top level root concept => +50

❖ Linkage axis (Horizontal relationships)

- Direct equivalence => 0
- Weak link => +5
- Strong link => +1

Based on the literature we opted for weighted edge counting as opposed to the simple edge counting for more accurate results [12]. As such, the rules for the Inheritance axis and the Linkage axis assign weights to edges. That the ontological distance vertically and horizontally between two concepts is a measure of their semantic relatedness was demonstrated by Rada [11] and is the basis of the *Reach Score*. Of the rules of particular note is the rule for the root concept. Consistent with the finding of Li [12], as the root concepts are the highest level of abstraction they must have a high weighting to correctly take into account that abstraction. Without this higher weighting, erroneous scores will result. For example, Method Acting is a child of Acting and Acting in turn is the child of Action. Moving up the tree from Method Acting to Acting yields a *Reach Score* of 1. However, moving up from Acting to action yields a *Reach Score* of 50, this reflects the correctly the degree of relationship between the concepts. Method Acting is a subset of Acting, where Acting is a concept with specific meaning. However, while Acting is a subset of action, action itself is a broad concept that without a higher weight would make Acting too close to other forms of actions such as Editing. Likewise, horizontally a lower weight is assigned to stronger links while assigning higher score to weaker links. The temporal axis then is used to give the concepts a temporal weight, which is very important within the domain as all concepts in the domain exist in a temporal context of the Production Cycle and the Motion Picture Life Stages. In the next section we discuss the results obtained by application of the metric to pairs of concepts.

6. Results

Table 1 shows the results of the application of the metric to a range of concept pairs. In terms of spread, we consider a score of 0 to indicate that two concepts are in fact synonyms. It must be noted that this will only occur in the case of "direct equivalence". Direct equivalence is a special-link by which concepts are connected in the ontology that effectively denotes one of the concepts to be a synonym of the other. An example that can be seen in the table is the concepts of Frame Rate Per Second and FPS, the latter being an abbreviation of the former. The reason that *Temporal Score* does not apply in this case, as per the rules, is because synonym concepts are not linked to the production cycle or the life stage on their own. They inherit these associations through their equivalency link with the central concept.

	Concept 1	Concept 2	Temporal Score	Reach Score	Total
1	Method Acting	Camera	2	117	119
2	Method Acting	Actor	1	2	3
3	Cross-cutting	Editor	1	4	5
4	Cross-cutting	Actor	3	116	119
5	Cross-cutting	Motion Picture	2	56	58
6	Motion Picture	Cinema	2	1	3
7	Motion Picture	Actor	1	3	4
8	Motion Picture	Editor	1	3	4
9	Treatment	Producer	1	1	2
10	Screening copy	Publicity	2	6	8
11	Screening copy	Prestige	6	12	18
12	Script	Camera	2	116	118
13	Script	Prop	2	53	55
14	Camera	Prop	2	116	118
15	Camera	Lighting	2	5	7
16	Frame rate per second	Fps	N/A	0	0
17	Mood	Genre	2	7	9
18	Mood	Tone	N/A	0	0
19	Mood	Category	2	4	6
20	Mood	Rating	2	8	10
21	Mood	Style	2	4	6
22	Film Festival	Prestige	4	13	17
23	Film Festival	Cinema	2	1	3
24	Film Festival	Award	4	5	9
25	Film Festival	Producer	1	65	66

Table 1: Semantic relatedness scores for pairs of Motion Picture Industry concept

A score of 1 to 10 is considered very close while scores above 100 firmly put two concepts in the "far" category and indeed the link between them is probably just through the top level root concepts. Score of 10 to 100 then represents a progression, with a score of 50 indicated that two concepts are through one or more concepts and may in fact have hit one of the root concepts.

As an example, Script and Prop yield a total score of 55 because both are instances of Artifact and are therefore connect through that root concept. Logically, the distance of 55 also makes sense because props are based on scripts. If the script calls for pirates, props designed for ninjas will not do. Also a script might to give description of props, i.e. the script might demand that the pirate captain have but one eye, thus dictate the use of an eye-patch prop but unless significant to the plotline the number and style of the pirate captains weapons and other props would be left to the discretion of the director. As such, 55 is an apt score for the semantic relatedness of script and prop.

An example of a higher score is Method Acting and Camera, the semantic relatedness score of this concept pair is 119. This is an apt score because these two concepts are only linked to each other through the root concepts of each, Action for Method Acting and Tool for Camera. Logically, this also makes sense given that while the camera would be filming the acting of an actor employing method acting skills, the camera does not have any tangible relationship with method acting beyond both being part of the Motion Picture Industry discourse.

The concept pairs that yield low scores are obviously related, e.g. Method Acting and Actor. However, scoring system also proved accurate when yielding mid-ranged score for concept pairs. A good example of this is screening copy and prestige. While the link between screening copy and publicity is obvious, the showing of the screening copy leads to publicity, the link between screening copy and prestige is more tenuous. There is a link and it is not too distant but neither is it a close relationship that can be easily defined. We believe, a score of 18 reflects the relatedness appropriately.

An unexpected benefit of the metric system turned out to be how it can work as a guide and a check. Anomalies in the scoring were found to be symptoms of faults in the ontology. As such, as the ontology is developed and refined further, the metric can serve as an error checking and completeness mechanism.

7. Future Work

The chief work yet to be undertaken in terms of the metric is a comprehensive evaluation against relatedness scores obtained from industry practitioners. We will be undertaking this evaluation with our industry partners AFTRS [13] in the very near future. In addition, we are currently in the process of integrating the metric into our information management system to assist with intelligent querying and data classification aspects. In turn, this aides us in our endeavors to exploit the Loculus ontology in order to support reuse and repurposing of data and to facilitate improvements to the production processes. We are also determining how to employ our method to calculate the semantic relatedness score for concept triples.

8. Conclusion

In this paper we presented a rule-based metric for calculating semantic relatedness score for the Motion Picture Industry. The metric is based on three axes: the *Temporal Axis*, the *Inheritance Axis* and the *Linkage Axis*. These three axes yield two scores: the *Temporal Score* and the *Reach Score*. The combined total of *Temporal* and *Reach Score* gives us the overall relatedness score for two concepts. The results to date are consistent with human expectations. Once integrated within our information system, we hope that the metric will aid us in the exploitation of the Loculus ontology to support reuse and repurposing of data and to facilitate improvements to the production processes of the Motion Picture Industry.

9. References

1. Ouyang, C., et al., *Camera, Set, Action: Automating Film Production via Business Process Management*, in *Creating Value: Between Commerce and Commons - CCI International Conference*. 2008: Brisbane Australia.
2. Ouyang, C., et al., *Towards Web-Scale Workflows for Film Production*, in *QUT ePrints Archive*. 2008.
3. Leigh, A., *Context! Context! Context!: Describing Moving Images at the Collection Level.* The Moving Image, 2006. **6**(1): p. 33-65
4. Tsinaraki, C., et al., *Ontology-Based Semantic Indexing for MPEG-7 and TV-Anytime Audiovisual Content.* Multimedia Tools and Applications, 2005. **26**: p. 299-325.
5. Lagoze, C. and J. Hunter, *The ABC Ontology and Model.* Journal of Digital Information, 2001.
6. Hunter, J., *Enhancing the Semantic Interoperability of Multimedia Through a Core Ontology.* IEEE Transactions on Cirsuits and Systems for video technology, 2003. **13**(1).
7. Budanitsky, A. and G. Hirst. *Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures*. in *Workshop on WordNet and Other Lexical Resources*. 2001. Pittsburgh, Pennsylvania, USA.
8. Blanchard, E., et al. *A typology of ontology-based semantic measures*. in *Open Interop Workshop on Enterprise Modelling and Ontologies for Interoperability* 2005. 2005.
9. Hirst, G. and D. St-Onge., *Lexical chains as representations of context for the detection and correction of malapropisms*, in *Fellbaum 1998*. 1998. p. 305–332.
10. Leacock, C. and M. Chodorow, *Combining local context and WordNet similarity for word sense identification*, in *Fellbaum 1998*. 1998. p. 265–283.
11. Rada, R., et al., *Development and application of a metric on semantic nets.* IEEE Transactions on Systems, Man and Cybernetics, 1989. **19**(1): p. 17-30.
12. Li, Y., Z.A. Bandar, and D. McLean, *An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources.* IEEE Transactions on Knowledge and Data Engineering, 2003. **15**(4): p. 871-882.
13. AFTRS. *AFTRS Homepage*. [cited; Available from: http://www.aftrs.edu.au/.